On the Relative Advantage of Mixed-Mode versus Single-Mode Surveys

Jorre T. A. Vannieuwenhuyze

Institute for Social & Economic Research, University of Essex, Colchester, UK

Survey researchers increasingly use mixed-mode surveys for general population data collection because mixed-mode surveys are argued to provide lower selection error at constant budgets, or lower variable costs at constant selection error. Nevertheless, the advantage of lower selection error and variable costs might be counteracted by higher measurement error and higher fixed costs. This trade-off between selection error, measurement error, variable costs, and fixed costs has hardly been studied within the existing literature about mixed-mode surveys. A possible procedure for evaluating this trade-off is discussed in this paper by comparing the performance (mean squared error) of mixed-mode survey designs against single-mode survey designs. The procedure is further illustrated by real example data stemming from a mixed-mode mailface-to-face survey. This illustration yields smaller errors for single-mode designs under low budgets but smaller errors for mixed-mode designs under large budgets or, alternatively, a budgetary advantage of single-mode designs when the allowed error is relatively high but a budgetary advantage of mixed-mode designs when the allowed error is relatively small. However, the validity of these results depend on several modeling assumptions which may be topics for future research.

Keywords: Mixed-mode surveys, Single-mode surveys, Mean Squared Error, Survey Design, Survey Budget, Survey Costs

1 Introduction

Traditionally, survey designs start from the choice of one particular data-collection mode like Computer-Assisted Personal Interviewing (CAPI), Computer-Assisted Telephone Interviewing (CATI), Mail Self-Administration Questionnaires (MSAQ's), or Web Self-Administration Questionnaires (WSAQ's). The choice of an optimal mode involves a trade-off between, among others, different forms of selection error and costs. Selection error is the error introduced by only observing a small subset of population members instead of the entire population because of noncoverage, sampling, or nonresponse. In general, cheap data-collection modes with low random selection error (like sampling error) go with high systematic selection error (like nonresponse or coverage error) and vice versa. For example, WSAQ surveys involve low development and implementation costs and thus allow for large sample sizes and low sampling error given the available budget. Nonetheless, Web surveys often suffer from much coverage and nonresponse error. CAPI surveys, in turn, are expensive and thus restrict the maximal possible sample size, but they are known to obtain less biased samples of population members.

In order to avoid a difficult choice between data-collection modes, survey methodologists have suggested using mixed-mode surveys instead of single-mode surveys (de Leeuw, 2005; Dillman, Smyth, & Christian, 2009). Mixed-mode surveys are surveys where different groups of respondents complete the survey by different data-collection modes. As such, they are argued to reduce selection error or costs relative to a single-mode survey, say by mode m_1, for two reasons. Firstly, if m_1 is a cheap mode with low random but high systematic selection error, certain population members might not be willing or able to respond to mode m_1 in the single-mode survey but do respond to another mode in the mixed-mode survey. In this case, the mixed-mode survey offers greater external validity than the single-mode survey. Secondly, if m_1 is an expensive mode with low systematic but high random selection error because of limited possible sample sizes, some respondents may respond by a cheaper mode in the mixed-mode survey so that larger samples can be drawn or total survey costs can be reduced. In this case, the mixed-mode survey offers greater external reliability than the single-mode survey.

Both reasons mentioned above make clear that mixed-mode surveys are only advantageous over single-mode surveys if *selection effects* occur (Vannieuwenhuyze & Loosveldt, 2013). Selection effects refer to differences between the respondents allocated to the different modes in a mixed-mode survey. If such selection effects are absent, then a single-mode design would provide samples which equally represent the population and a mixed-mode design would thus be worthless to use. Nevertheless, the advantage of selection effects within mixed-mode surveys might not be guar-

Contact information: Jorre T. A. Vannieuwenhuyze, Institute for Social & Economic Research, University of Essex, Colchester, UK, email: jtavan@essex.ac.uk

anteed because of two reasons.

First, the use of a mixed-mode instead of a single-mode design requires additional fixed costs for the development and implementation of the additional modes. These additional fixed costs might overwhelm a reduction in variable costs per survey respondent. The choice between mixed-mode and single-mode designs thus firstly involves a trade-off between fixed and variable costs.

Second, selection effects might be counteracted by other types of mode effects, namely *measurement effects* (de Leeuw, 2005; Voogt & Saris, 2005; Dillman, Smyth, & Christian, 2009; Weisberg, 2005). Measurement effects are differences in measurement error accompanying the different data-collection modes (Voogt & Saris, 2005; Weisberg, 2005) and occur when answers of the same respondents would differ if different data-collection modes were used. Typical examples of measurement effects are social desirability error in interview modes, recency effects in telephone surveys, or primacy effects in self-administration modes (see, among others, de Leeuw, 2005; Dillman, Smyth, & Christian, 2009; Schwarz, Strack, Hippler, & Bishop, 1991). The choice between mixed-mode and single-mode designs thus secondly involves a trade-off between selection error and measurement error.

In short, error and cost reduction by mixed-mode designs due to selection effects may be counteracted by additional fixed costs and by the occurrence of measurement effects. As a consequence, the advantage of mixed-mode surveys relative to single-mode designs is not automatically guaranteed. So far, the trade-off between both sorts of design has hardly been a topic of study. This paper aims to fill this gap by discussing possible procedures to evaluate a particular mixed-mode survey design relative to its single-mode counterparts.

The paper is further organized as follows. Section 2 introduces and discusses the Mean Squared Error (MSE) as a quantity to measure the performance of survey designs under a given budget. The MSE of two distinct designs can subsequently be compared in order to determine the advantage of one design relative to the other design. Subsequently, Section 3 provides an illustration of the procedure using a survey item about opinions about surveys. In this illustration, a single-mode mail design, a single-mode face-to-face design, and a sequential mixed-mode mail—face-to-face design are compared. Finally, Section 4 concludes the paper with a short repetition and discussion of the drawbacks of the proposed procedure.

2 A procedure to evaluate mixed-mode designs

This section sketches a possible procedure to evaluate the usefulness of a particular mixed-mode survey design compared to one of its single-mode counterparts when real data is available from both these mixed-mode and single-mode designs. For simplicity, let us restrict focus to the estimation of the mean $\mu = \mu(Y)$ of a target variable Y under simple random sampling, and let us restrict the number of modes to two, i.e. modes m_1 and m_2. Extensions to situations

with other statistics, other sampling methods, and more than two modes can be derived from the following discussion but might require more complex analysis frameworks.

A frequently used quantity to compare the quality of survey designs is the *Mean Squared Error* (MSE). The MSE of a sample mean $\bar{Y}_{\delta,\beta}$ obtained by a survey using design δ within a budget constraint β is defined as

$$\text{MSE}(\bar{Y}_{\delta,\beta}) = E[(\mu - \bar{Y}_{\delta,\beta})^2]. \tag{1}$$

The budget constraint β refers to the total available amount of money which the researchers spend on the entire survey process. The survey design δ, in turn, includes all other decisions to be made by the researchers, also including the cost allocation of the total budget to different aspects of the survey process.

The researchers' task now is to select the optimal survey design which either minimizes the MSE given an available budget or minimizes the required budget given a maximal allowable MSE. Put differently, within the context of mixed-mode surveys, researchers need to compare $\text{MSE}(\bar{Y}_{mm,\beta})$ obtained by a mixed-mode design $\delta = mm$ with $\text{MSE}(\bar{Y}_{sm,\beta})$ obtained by a single-mode design $\delta = sm$ using, for example, mode m_1. This comparison can be quantified by the difference between both MSE's, that is $\text{MSE}(\bar{Y}_{mm,\beta}) - \text{MSE}(\bar{Y}_{sm,\beta})$.

The MSE in (1) can further be decomposed into

$$\text{MSE}(\bar{Y}_{\delta,\beta}) = [\mu - E(\bar{Y}_{\delta,\beta})]^2 + \text{Var}(\bar{Y}_{\delta,\beta}), \tag{2}$$

where $[\mu - E(\bar{Y}_{\delta,\beta})]$ is the systematic error or bias, and $\text{Var}(\bar{Y}_{\delta,\beta})$ the random error of $\bar{Y}_{\delta,\beta}$. Under simple random sampling, $E(\bar{Y}_{\delta,\beta})$ is known to equal $E(Y_{\delta,\beta})$, where $Y_{\delta,\beta}$ represents the outcome variable in surveys using design δ and budget β. Likewise, under simple random sampling, $\text{Var}(\bar{Y}_{\delta,\beta})$ is known to equal $\text{Var}(Y_{\delta,\beta})/N_{\delta,\beta}$, where $N_{\delta,\beta}$ denotes the net sample size of a survey using design δ and budget β. However, the outcome variable $Y_{\delta,\beta}$ does not depend on the survey budget β because, by definition, the measurement of this variable is solely determined by survey design δ. In other words, subscript β can be dropped from $Y_{\delta,\beta}$. Taking all together, the MSE can be written as

$$\text{MSE}(\bar{Y}_{\delta,\beta}) = [\mu - E(Y_\delta)]^2 + \frac{\text{Var}(Y_\delta)}{N_{\delta,\beta}}. \tag{3}$$

According to (3), estimation of a difference in MSE's requires knowledge of, firstly, $E(Y_\delta)$ and $\text{Var}(Y_\delta)$ for both the mixed-mode and single-mode designs, secondly, $N_{\delta,\beta}$ for both the mixed-mode and single-mode designs, and, thirdly, the target mean μ. Convenient estimates of $E(Y_\delta)$ and $\text{Var}(Y_\delta)$ are the sample mean $\bar{y}_\delta$ and sample variance s_δ^2 of Y obtained from any survey using design δ. This sample mean and variance can straightforwardly be implemented in (3). Estimation of the net sample size and the target mean requires more elaborate discussion as provided in the next subsections.

Estimation of the net sample size

Unlike the sample mean and the sample variance, the net sample size does depend on the survey budget β and creates

the link between the MSE and the budget. As a consequence, it should be expressed as a function of β. Consequently, either the budget β can be fixed to a specific value and the MSE of the mixed-mode and single-mode designs can be calculated and compared given this specified budget, or the MSE can be fixed to a specific value and the required budget of the mixed-mode and single-mode designs can be calculated and compared given this specified MSE.

The discussion in this section starts from a scenario where gross sample size and budget follow a linear relation and where fixed costs and variable costs per drawn population member ($\neq$respondent) are known for all survey designs. This scenario overlaps with the available knowledge of the illustration in the next section. However, in other situations, the relation between gross sample size and budget might not be linear and the available knowledge about costs might be different (for example, one might know the variable cost per respondent instead of per drawn sample member). In these situations, the formulas outlined in this section should be adapted, even though the main idea of writing the net sample size in terms of the survey budget remains the same.

The net sample size is not constant but variable. Indeed, people with response propensities smaller than 1 may cause differences in net sample sizes over different replications of a survey using the very same design δ and the very same budget β. The gross sample size (i.e. the total sample randomly drawn from the sampling frame), in contrast, can be treated as fixed given a specific design δ and budget β. The net sample size relates to the gross sample size $M_{\delta,\beta}$ by the equation

$$N_{\delta,\beta} = M_{\delta,\beta} P_\delta, \tag{4}$$

where P_δ is the response rate calculated on the total gross sample size (i.e. including ineligible sample members). A convenient estimate of this response rate is the sample response rate p_δ obtained from any survey using design δ. Like $Y_{\delta,\beta}$, it is straightforward to assume that this response rate only depends on the survey design and not on the budget β.

The next question is how the gross sample size relates to the survey budget. When the relationship between gross sample size and budget is linear, the total survey budget factorizes into

$$\beta = f_\delta + M_{\delta,\beta} v_\delta, \tag{5}$$

where f_δ refers to the fixed survey costs of a survey using design δ, and v_δ refers to the variable survey costs per drawn sample member of a survey using design δ. The fixed costs include, among others, the survey organization, the development and implementation of the questionnaire, and logistics. The variable costs include, among others, the cost for interviews, questionnaire sending, and quality control.

Substituting (4) into (5) and after some rearrangements, we get

$$N_{\delta,\beta} = \frac{(\beta - f_\delta)P_\delta}{v_\delta}. \tag{6}$$

This equation allows for estimation of the net sample size and can be implemented in (3), provided that the fixed and variable survey costs for design δ are known.

Estimation of the target mean

Absolute estimation of the target mean μ is a hard (if not impossible) task, because it requires knowledge of the target variable within the entire target population and without measurement error. Observed survey data rarely allow for the estimation of the target mean because of selection and measurement error. An alternative is to estimate a relative target mean starting from a benchmark data-collection mode and a reference sample. This alternative boils down to the next two assumptions:

Assumption 1: Ignorable measurement error of the benchmark mode. Usually, the target variable Y is defined as a theoretical construct. However, by definition, such a theoretical construct excludes measurement and is useless in practical evaluation of survey error. An alternative is to use a variable measured by a particular mode, say mode m_1, as a benchmark and to compare the outcomes of other modes with this variable. It is thus assumed that the benchmark mode m_1 comes with negligible measurement error while mode m_2 is considered a distorting mode. Put differently, measurement error of mode m_1 is considered ignorable and measurement by mode m_2 is evaluated relative to measurement by mode m_1.

Assumption 2: Ignorable selection error in the reference sample. Ideally, the target mean is measured on the entire survey population. However, such measurement is infeasible in practice because of sampling, noncoverage, and nonresponse. An alternative is to select one realized survey sample as a reference population. This selected sample must represent the target population in the best way according to our beliefs. It is thus assumed that the reference sample comes with negligible selection error while other samples are considered distorted samples. Put differently, selection error in this reference sample is considered ignorable and other samples are evaluated relative to this reference sample.

Because mixed-mode surveys are generally used to lower selection error relative to their single-mode counterparts (de Leeuw, 2005; Dillman, Smyth, & Christian, 2009), the most logical candidate reference sample is the sample obtained from the mixed-mode survey. Indeed, mixed-mode survey designs are introduced by the belief that they provide samples which, on average, better represent the population compared to their single-mode survey counterparts. Nevertheless, in situations where this proposition does not hold, a single-mode sample can be taken as the reference population instead.

However, choosing one single mode m_1 as benchmark mode and a mixed-mode dataset as reference sample creates a problem. By definition, part of the reference sample data is not collected by the benchmark mode m_1 but by the distorted mode m_2, and this mismatch prevents direct estimation of the target mean. Indeed, estimation of the target mean requires all data to be collected by mode m_1, including the data of

respondents who actually respond by mode m_2. The question is how mode m_2 respondents would respond if their data had been collected by mode m_1 instead of mode m_2. Such data are called *counterfactuals* or *potential outcomes* (Rubin, 1974; Rosenbaum & Rubin, 1983; Holland, 1988; Galles & Pearl, 1998).

The causal inference literature (e.g., among others, Pearl, 2009; Morgan & Winship, 2009; Weisberg, 2010) provides two general analysis models to avoid counterfactual data and to estimate the target mean within mixed-mode data. These models start from conditioning the analysis model on well-chosen covariates and are briefly described below. A detailed general discussion of both methods is given by Pearl (1995) and Pearl (2000), and the application of both methods within the context of mixed-mode data is discussed in detail by Vannieuwenhuyze et al. (2014).

The back-door model. The first and most popular model for estimating counterfactuals involves the inclusion of covariates B, where B is argued to explain why different respondents are selected for different modes. This model is called the *back-door model* (Pearl, 1995, 2009), because it aims to capture 'back-door' correlations between the survey mode and the target variable which arise from common cause variables. Within the mixed-mode literature, the back-door model has already been widely applied, but predominantly using socio-demographic variables as back-door covariates (e.g., among others, Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011; Heerwegh & Loosveldt, 2011; Jäckle, Roberts, & Lynn, 2010; Hayashi, 2007; Fricker, Galesic, Tourangeau, & Yan, 2005; Holbrook, Green, & Krosnick, 2003; Greenfield, Midanik, & Rogers, 2000).

However, a covariate B must meet two important assumptions in order to consistently estimate the target mean (Pearl, 2009; Morgan & Winship, 2009). The first assumption is that B fully captures the selection effect between the modes (Rosenbaum & Rubin, 1983), and this assumption is called the ignorable mode allocation assumption because it refers to a random allocation of the respondents to both modes after control for B. The second assumption is the absence of measurement effects on B (Pearl, 2009), and this assumption is called the mode-insensitivity assumption because it refers to the B variables being mode-insensitive.

If both the ignorable mode selection assumption and the mode-insensitivity assumption hold true, it can be shown that the target mean can be written as an expression of quantities which do not require counterfactual data (Vannieuwenhuyze, Loosveldt, & Molenberghs, 2014). If B is a discrete variable, then the following result emerges:

$$\mu = \sum_b \mu_{m_1 b} \left(\pi_{b|m_1} \tau_{m_1} + \pi_{b|m_2} \tau_{m_2} \right), \qquad (7)$$

where $\mu_{m_1 b}$ represents the conditional mean $E(Y|B = b)$ within the mode m_1 group of respondents, $\pi_{b|m}$ represents the probability $P(B = b)$ within the mode m group of respondents, and τ_m represents the proportion of respondents selected for mode m. The right-hand side of (7) can be estimated from observed mixed-mode data and can be imple-

mented in (3). As a result, if a covariate B is available which satisfies both assumptions, the back-door model allows for evaluating the MSE of different survey designs.

The front-door model. The second model for estimating counterfactuals involves the inclusion of a covariate F, where F explains differences in measurement error between the different modes. This model is called the *front-door model* (Pearl, 1995, 2009) because it aims to capture 'front-door' correlations between the survey mode and the target variable which arise from a direct causal effect of the survey mode on the variable of interest. Nonetheless, the front-door model remains relatively unexplored within the current mixed-mode literature. This model is solely used by Vannieuwenhuyze et al. (2014) who use a question about survey pleasure as the front-door covariate.

However, like back-door covariates, covariate F must meet two assumptions in order to consistently estimate the target mean (Pearl, 2009; Morgan & Winship, 2009). The first assumption is that F fully captures the measurement effect between the modes and this assumption is called the exhaustiveness assumption. The second assumption is the absence of selection effects on F and this assumption is called the isolation assumption because F must be isolated from selection effects.

Like with the back-door model, if both the exhaustiveness assumption and the isolation assumption hold true, it can be shown that the target mean can be written as an expression of quantities which do not require counterfactual data (Vannieuwenhuyze et al., 2014). If F is a discrete variable, then the following result emerges:

$$\mu = \sum_f \pi_{f|m_1} \left(\mu_{m_1 f} \tau_{m_1} + \mu_{m_2 f} \tau_{m_2} \right), \qquad (8)$$

where $\pi_{f|m_1}$ represents the probability $P(F = f)$ within the mode m_1 group of respondents, and μ_{mf} represents the conditional mean $E(Y|F = f)$ within the mode m group of respondents. The right-hand side of (8) can be estimated from observed mixed-mode data and can be implemented in (3). As a result, if a covariate F is available which satisfies both assumptions, the front-door model also allows for evaluating the MSE of different survey designs.

3 An illustration with the Survey On Surveys

This section provides an illustration of the technique outlined in the previous section. In this illustration, it will be investigated whether a sequential mixed-mode mail—face-to-face design provides better results than its single-mode mail and face-to-face counterpart designs for the estimation of the average opinion about surveys. The relative performance of all three designs is estimated by the Survey on Surveys data, which was organized in 2004 in Flanders, Belgium, by the Survey Methodology Research Group of the Centre for Sociological Research, KU Leuven (Storms & Loosveldt, 2005).

Survey design

The total sample of the Survey on Surveys consisted of 1200 Flemish persons aged between 18 and 80 sampled from the national register. A two-stage sampling procedure was used in which first 48 communities where drawn with a chance proportional to size and with replacement. Subsequently, 25 people were randomly drawn from each selected community. Because of sampling with replacement, 45 communities were effectively drawn of which one was drawn three times and one was drawn two times. These two communities thus include 75 and 50 sample members respectively. The clustering within communities will be taken into account in the analysis, and nonresponse is further assumed ignorable conditional on the communities.

Note that the sampling technique of the Survey On Surveys is not a simple random sample, as is used in the theoretical Section 2. However, it can easily be assumed that the sampling technique does not affect measurement error or that sample statistics from a simple random sample can straightforwardly be estimated by the clustered two-stage random sample of the Survey On Surveys. The only difference will be that larger standard errors for the sample statistics might be obtained under two-stage random sampling relative to simple random sampling. Nonetheless, the reader should bear in mind that the conclusions below only apply for survey designs using simple random sampling and not two-stage random sampling.

Each sample member of the Survey On Surveys was first randomly selected for one of two arms. The first arm included 20 sample members per selected community, while the second arm included 5 sample members per selected community. In sum, 960 sample members were assigned to the first arm and 240 were assigned to the second arm. A sample member of the first arm was first contacted by mail with an invitation to complete an enclosed paper questionnaire. If this sample member did not return the paper questionnaire, a first reminder was sent by mail two weeks later and a second reminder accompanied by a new questionnaire was sent by mail four weeks after the first reminder. The mail survey phase lasted two months. Sample members who did not return the paper questionnaire in due time were contacted by an interviewer at home to complete a face-to-face (ftf) interview (i.e. CAPI). Nonetheless, this face-to-face follow-up was unknown to the sample members during the initial mail phase. A sample member of the second arm was immediately contacted by an interviewer at home to complete a face-to-face interview. The survey questionnaire and data collection strategy of this second arm were equal to the second phase of the first arm, except that the sample members did not receive the mail questionnaire first. In both arms, a €5 gift voucher was used as an incentive for returning the questionnaire or for responding to the interview.

The data allow comparison of three survey designs. The first design is the sequential mixed-mode design from the first arm, starting with the mail phase and ending with the face-to-face phase, and we will further refer to this design by $\delta = $ mm. The second design is a single-mode mail design which can be evaluated by the responses of the first phase of the first arm, and we will further refer to this design by $\delta = $ mail. The third design is a single-mode face-to-face design which can be evaluated by the data from the second arm, and we will further refer to this design by $\delta = $ ftf. In summary, the data allow to compare a sequential mixed-mode design with its single-mode mail and single-mode face-to-face counterparts.

Because the aim of the paper is to illustrate the evaluation techniques rather than to make judgments about the population, the analysis will only include those respondents who responded to all the variables listed below. Partial responses are considered as nonresponse. If only full response is considered, the single-mode mail survey (i.e. the initial mail phase of the first arm) reached a response rate of 55.4% (Table 1), while the mixed-mode survey (i.e. the entire first arm) reached a response rate of 72.9%, which is a relatively high response rate for a general population survey. The single-mode face-to-face survey (i.e. the second arm), in turn, had a response rate of 71.7%. Note that the number of mail respondents in the single-mode mail design is lower than the number of mail respondents in the mixed-mode design. This difference is caused by the fact that some respondents sent back their mail questionnaire after a face-to-face contact in the follow-up phase. We assume that these people would not have responded if they were not recontacted face-to-face.

Variables

Target variable Y. The MSE is analysed on the mean of an item about the respondents' opinions about surveys (Loosveldt & Storms, 2008). Each respondent was asked to indicate (dis-)agreement with the proposition "I do not like participating in surveys" on a 5-point Likert-scale ranging over '1=completely disagree', '2=disagree', '3=neither agree nor disagree', '4=agree', and '5=completely agree'. In the mail questionnaire, these answer categories were listed horizontally in a table but a 'don't know'/'no opinion' option was not provided. In the face-to-face interviews, the response categories were read out by the interviewer and presented vertically on a showcard, again excluding 'don't know' and 'no opinion' options.

The particularity of this item about survey liking might very likely cause selection effects and measurement effects on the mean. First, there might be a selection effect within the sequential mixed-mode design as non-respondents to the mail questionnaire are likely to be more negative about surveys (Loosveldt & Storms, 2008). Second, a measurement effect is also expected because respondents interviewed face-to-face will probably tend to report more socially desirable positive opinions about surveys in front of an interviewer (Dillman, Phelps, et al., 2009; Loosveldt & Storms, 2008). It should be stressed, however, that this item about survey liking is strategically chosen for the purpose of the illustration, but the particularity of this item probably prevents drawing general conclusions from this illustration.

Front-door variable F. Within this illustration, the front-door model is used for the estimation of the target mean μ.

Table 1 The Survey On Surveys includes data collected by mail questionnaires and face-to-face (ftf) interviews and allows comparing three survey designs.

	mixed-mode mail-ftf design	single-mode mail design	single-mode ftf design
Frequency of . . .			
mail response	525	495	
face-to-face response	130		160
partial response	32	26	2
nonresponse	211	426	61
not eligible	62	13	17
Response rate*	0.729	0.554	0.717

* = response/(total - not eligible)

As a front-door variable, a question is used which asks for the respondents' experiences during the survey. At the end of the questionnaire, the respondents were asked whether they found answering the questions a pleasant or unpleasant task. The respondents could select an answer from a 5-point Likert-scale including the responses 'very pleasant', 'pleasant', 'neither pleasant, nor unpleasant', 'unpleasant', and 'very unpleasant'. The format of this question in the mail questionnaire and the face-to-face interview was exactly equal to the opinion about survey items. Because relatively few respondents marked 'very pleasant', 'unpleasant', and 'very unpleasant', the variable was dichotomized ('very pleasant' and 'pleasant' versus 'neither pleasant, nor unpleasant', 'unpleasant' and 'very unpleasant'). This question about survey pleasure is used as a front-door covariate because Vannieuwenhuyze et al. (2014) found that this covariate yielded mode effects estimates in line with expectations, while a back-door model with socio-demographic covariates did not. Unfortunately, other potential back-door or front-door covariates are not available in the Survey on Surveys. Developing and testing other potential covariates is a topic which certainly needs to be addressed in future research.

It is very likely that the mode of data-collection has a direct causal effect on reported survey pleasure. Indeed, survey pleasure may be higher in face-to-face surveys because respondents enjoy the interaction with an interviewer. Survey pleasure, in turn, probably has an effect on the reported opinion about surveys because people who report completing the survey as a pleasant task will tend to report more positive opinions about surveys in general.

Nevertheless, the central question remains whether survey pleasure fully captures the measurement effect on the target variable. Some insights can be given by the associations between the mode group and the front-door variable, and between the front-door variable and the target variable. There is a significant association between the mode group and survey pleasure ($\chi^2 = 34.431$, $df = 1$, $p < 0.001$). Moreover, even though the face-to-face mode includes more reluctant population members who did not complete the initial mail questionnaire, the face-to-face respondents report a significant higher pleasure compared to the mail respon-

dents. This observation may confirm that people enjoy the interaction with an interviewer. Likewise, the association between survey pleasure and the opinion item is positive and highly significant as well ($\chi^2 = 63.369$, $df = 4$, $p < 0.001$). This observation may confirm that people experiencing more pleasure in answering the survey will report a more positive opinion about surveys in general. Nevertheless, it should be noted that these correlations between the mode group, the front-door variable, and the target variable neither prove the presence of a measurement effect nor prove the capturing power of the front-door variable.

Costs

The Survey On Surveys was entirely collected by the Survey Methodology Research Group of the Centre for Sociological Research of the KU Leuven. As a consequence, it is hard to define the exact fixed and variable costs of all three survey designs. As a solution, a new price offer was asked to the Belgian commercial survey organization GfK Significant for all three designs. These designs included the very same questionnaire, organization, and implementation as the Survey On Surveys except that a simple random sample was used instead of a two-stage random sample. The total prices were asked for the three designs when the total gross sample would equal 1000, 1500, 2000, 2500, and 3000 people respectively.

For each design, the offered prices and gross sample sizes lie more or less on a straight line. As a consequence, these price offers allowed us to estimate the fixed and variable costs for each design by fitting a linear regression line through the five points for each design separately (Table 2).

The intercepts of these regression lines represent the fixed costs because they refer to the price of the surveys when the gross sample size is zero. The fixed costs of the single-mode mail survey add up to €14 456. As could be expected, the fixed costs of a single-mode face-to-face survey are somewhat larger, namely €20 591, because face-to-face surveys require expensive interviewer training. The fixed costs of the sequential mixed-mode mail-face-to-face survey only slightly exceed the fixed costs of a single-mode face-to-face survey, namely €21 574.

The slopes of the regression lines, in turn, denote the

Table 2 The three survey designs have different fixed and variable costs (in €).

	fixed cost f_δ	variable cost[*] v_δ
single-mode mail (δ = mail)	14 455.60	17.94
single-mode face-to-face (δ = ftf)	20 591.30	96.46
mixed-mode mail—face-to-face (δ = mm)	21 574.46	34.04

[*] *cost per randomly drawn population member*

variable costs as they represent the increase in total survey costs per added sample member. A mail survey has the lowest variable cost, which equals €18. A face-to-face survey, in contrast, involves large variable costs, namely €96, because interviewers need to be paid. The variable costs of a mixed-mode survey fall in between the above figures, i.e. €34 per sampled individual.

Estimation procedure

It is a complex task to estimate the difference in MSE between two designs because it is a non-linear function of different statistics including variance parameters. For that reason, a Bayesian analysis is used, in which the uncertainty about the population difference between two MSE's θ is formalized by the posterior probability distribution of θ given the available data, that is $p(\theta|\text{data})$. Using Bayes Rule, the posterior distribution is then calculated by

$$p(\theta|\text{data}) = c^{-1} \ p(\text{data}|\theta) \ p(\theta),$$

where c is a normalizing constant independent of θ, $p(data|\theta)$ is the likelihood function of the data for θ, and $p(\theta)$ is the prior distribution. The prior distribution reflects the researcher's uncertainty about θ before the data is taken into account, and, for that reason, may add a flavour of subjectivity to the analysis. Nonetheless, such subjectivity can easily be avoided if relatively flat priors are chosen.

A Bayesian analysis strategy has the advantage of computational ease combined with good frequentist properties for complex problems like ours. First, inferences about θ can easily be made in a Bayesian analysis by merely taking a number of random draws from its posterior distribution (Gelman, Carlin, Stern, & Rubin, 2004). The sample variance of these random draws provide estimates of uncertainty that are usually easier to calculate than asymptotic variances because random draws of θ are simple functions of random draws of the separate sufficient statistics for θ, that are all quantities in (3), (6), and (8) (Little, 2013). Second, in complex problems, Bayesian methods with relatively flat priors have been shown to yield better frequentist properties than many regular frequentist approaches because they automatically incorporate non-normality of sample variances and corrections for finite sample sizes (Little, 2004, 2012). Nonetheless, in order to have good frequentist properties, the analysis models should incorporate key design features through weights and clustering. For that reason, weighted parameters are estimated by using hierarchical models that take into account the clustered nature of the data within the communities. Let us further label the communities by the indices $c = 1, \ldots, 45$, and let us label the respondents by the indices $i = 1, \ldots, n_c$, where n_c denotes the net sample size within community c.

The design-specific means ($E(Y_\delta)$) and variances (Var(Y_δ)) are estimated by the hierarchical normal model:

$$y_{ic} \sim N(\theta_c, \sigma^2)$$
$$\theta_c \sim N(\theta, \tau^2),$$

where θ has been assigned a normal prior distribution $N(2.5, 10^6)$, and σ^2 and τ^2 have been assigned a half normal prior distribution $N(0, 10^6)I_{[0,\infty[}$. The mean is subsequently estimated by the weighted sum $\sum_{c=1}^{45} \theta_c \cdot M_c/M$ where M_c denotes the gross sample size of cluster c and M denotes the overall gross sample size. The variance is estimated by the sum $\sigma^2 + \tau^2$. The design-specific proportions (P_δ) are estimated by the hierarchical logistic model

$$y_{ic} \sim b(\pi_c)$$
$$\text{Logit}(\pi_c) \sim N(\lambda, \varsigma^2),$$

where λ has been assigned a normal prior distribution $N(0, 10^6)$, and ς^2 has been assigned a half normal prior distribution $N(0, 10^6)I_{[0,\infty[}$. The proportion is subsequently estimated by the weighted sum $\sum_{c=1}^{45} \pi_c \cdot M_c/M$.

In order to make inferences, Gibbs sampling was used via the WinBugs software (Lunn, Thomas, Best, & Spiegelhalter, 2000). One thousand simulated values of the sample means, variances, and proportions were taken from their posterior distributions by using every 100th value of 100 000 simulations after a burn-in of 1 000 iterations. Afterwards, these 1 000 values are used to calculate the differences between the MSE's as defined in (3). Point estimates are obtained by taking the median of these 1 000 simulated MSE-scores and 95% credibility intervals (CI's) are obtained by using the 2.5th and 97.5th percentiles of these 1 000 simulated MSE-scores.

Results

This section compares and discusses the estimated MSE's of the three survey designs, i.e. the single-mode mail survey (δ = mail), the single-mode face-to-face survey (δ = ftf), and the sequential mixed-mode mail—face-to-face survey (δ = mm). The mail questionnaire as well as the face-to-face interview are successively taken as the benchmark mode, which is the mode which goes without measurement error (see Assumption 1).
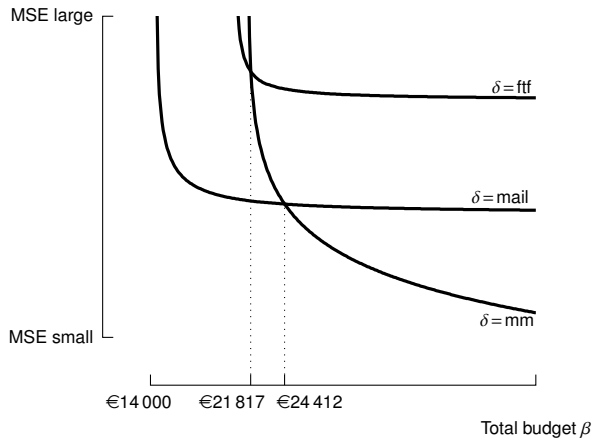
Figure 1. When the mail questionnaire is the benchmark mode and a large budget (β) is available, the MSE of the mixed-mode mail – face-to-face design ($\delta = $ mm) drops below the MSE's of the single-mode face-to-face design ($\delta = $ ftf) and the single-mode mail design ($\delta = $ mail).

Mail questionnaire as benchmark mode. When the mail questionnaire is chosen as the benchmark mode, the first surprising result is that the MSE of the single-mode face-to-face survey is always larger than the MSE of the single-mode mail survey (Figure 1). These results might point at an absolute advantage of using a mail survey instead of a face-to-face survey under any available budget. The second surprising result is that the MSE of the mixed-mode design has a steep drop at small budgets and immediately drops below the MSE curves of the single-mode designs. This means that the mixed-mode design seems to be more attractive than the single-mode designs as soon as the fixed costs of the mixed-mode design are covered by the total budget. Put differently, when the maximal allowable MSE is relatively large, a single-mode mail survey is cheaper, but when the maximal allowable MSE is relatively small, the mixed-mode design becomes more attractive.

The difference between the MSE's of the mail survey and the mixed-mode survey (Figure 2) crosses zero at a budget of €24 412 (With 95%-CI = [€23 127, €27 651]). This means that, when the available survey budget is lower than about €24 412, a single-mode mail survey will provide less error relative to the mixed-mode mail—face-to-face survey. When the available budget exceeds €24 412, in contrast, the mixed-mode design is more attractive. Note that, when the fixed and variable costs of the mail survey are taken into account, a budget of €24 412 implies a maximal gross sample of 555 people ($=(€24 412 - f_{mail})/v_{mail}$) which is a rather small sample size.

The difference between the MSE's of the face-to-face survey and the mixed-mode survey (Figure 3) crosses zero at a budget of €21 817 (With 95%-CI = [€21 723, €22 012]). This means that, when the available survey budget is lower than €21 817, a single-mode face-to-face survey will provide less error relative to the mixed-mode mail—face-to-face sur-
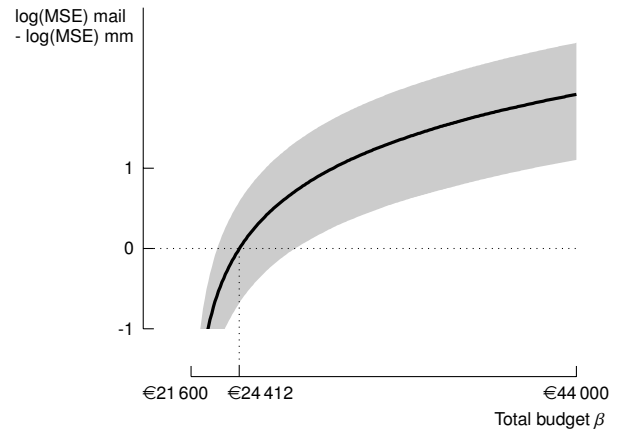

Figure 2. When the mail questionnaire is the benchmark mode, the difference between the MSE's of the mail survey and the mixed-mode survey crosses zero at a budget of €24 412 with a 95% credibility interval of [€23 127, €27 651].
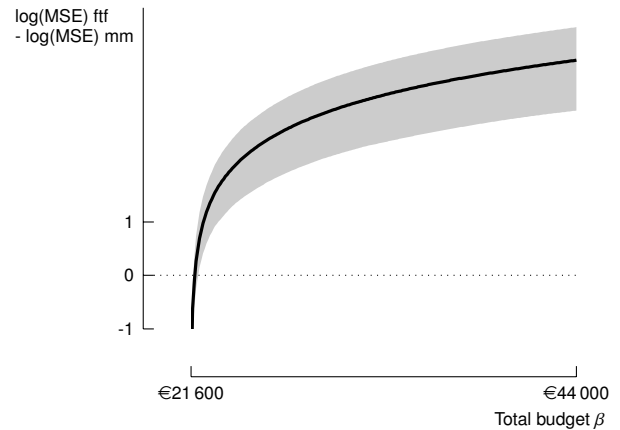

Figure 3. When the mail questionnaire is the benchmark mode, the difference between the MSE's of the face-to-face survey and the mixed-mode survey crosses zero at a budget of €21 817 with a 95% credibility interval of [€21 723, €22 012].

vey and vice versa. However, when the fixed and variable costs of the face-to-face survey are taken into account, a budget of €21 817 implies a maximal gross sample of merely 12 people. This sample size is unrealistically small and we can conclude that the mixed-mode design generally performs better than the single-mode face-to-face design in this situation.

Face-to-face interview as benchmark mode. Like in the previous subsection, when the face-to-face interview is chosen as the benchmark mode, the MSE of the single-mode face-to-face survey is always larger than the MSE of the single-mode mail survey (Figure 4). This result is even more surprising because the face-to-face survey is considered the unbiased benchmark mode here. Apparently, the large variable costs (or sampling error) of the face-to-face survey is not counterbalanced by a lack of measurement bias. Like in
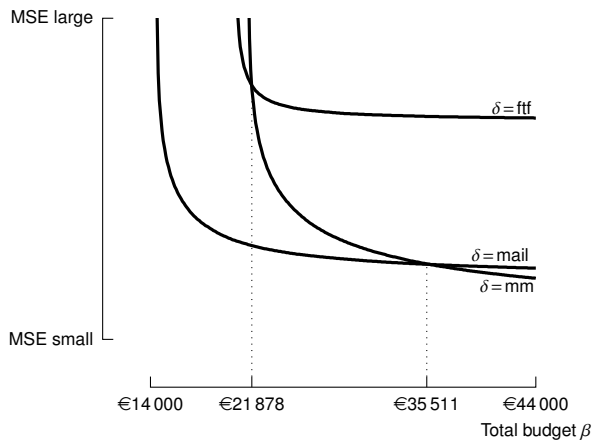
Figure 4. When the face-to-face interview is the benchmark mode and a large budget (β) is available, the MSE of the mixed-mode mail—face-to-face design ($\delta = $ mm) drops below the MSE's of the single-mode face-to-face design ($\delta = $ ftf) and the single-mode mail design ($\delta = $ mail).

the previous subsection, the MSE of the mixed-mode design has a steep drop at small budgets and immediately crosses the MSE of the face-to-face design. The curve of the mail survey MSE is, however, crossed at a larger budget. Put differently, like in the previous subsection, when the maximal allowable MSE is relatively large, a single-mode mail survey is cheaper, but when the maximal allowable MSE is relatively small, the mixed-mode design becomes more attractive even though it requires a large budget.

The difference between the MSE's of the mail survey and the mixed-mode survey (Figure 5) crosses zero at a budget of €35 511 (With 95%-CI = [€24 235,∞[). This would mean that the mixed-mode design becomes advantageous over the single-mode mail design whenever the available budget exceeds €35 511. Note however that the 95% credibility interval lower bound does not cross zero. So, there is only weak evidence of the advantage of the mixed-mode design for budgets larger than €35 511. When the fixed and variable costs of the mail survey are taken into account, a budget of €35 511 implies a maximal gross sample of 1173 people, which is a fair sample size.

The difference between the MSE's of the face-to-face survey and the mixed-mode survey (Figure 6) crosses zero at a budget of €21 878 (With 95%-CI = [€21 753,€22 174]). This means that, when the available survey budget is lower than €21 878, a single-mode face-to-face survey will provide less error relative to the mixed-mode mail—face-to-face survey and vice versa. However, when the fixed and variable costs of the face-to-face survey are taken into account, a budget of €21 878 implies a maximal gross sample of only 13 people. Like in the previous subsection, this sample size is unrealistically small and we can conclude that the mixed-mode design generally performs better than the single-mode face-to-face design in this situation.
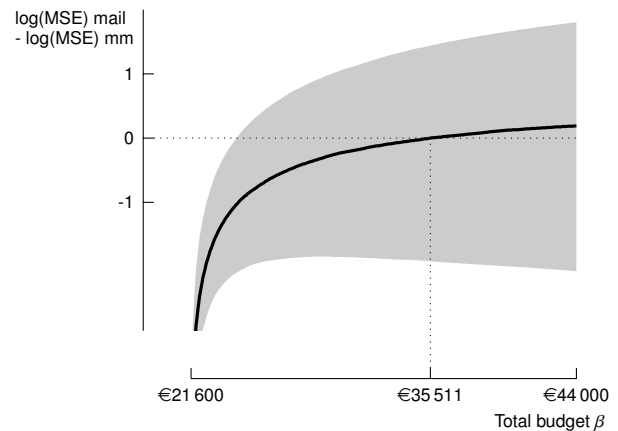

Figure 5. When the face-to-face interview is the benchmark mode, the difference between the MSE's of the mail survey and the mixed-mode survey crosses zero at a budget of €35 511 with a 95% credibility interval of [€24 235,∞[.
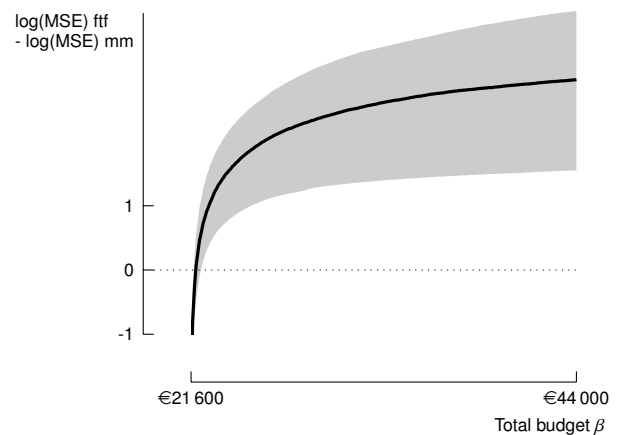

Figure 6. When the face-to-face interview is the benchmark mode, the difference between the MSE's of the face-to-face survey and the mixed-mode survey crosses zero at a budget of €21 878 with a 95% credibility interval of [€21 753,€22 174].

Summary and limitations of the illustration

The conclusion of this illustration would be that, in most situations, a mixed-mode mail—face-to-face design is more attractive than a single-mode mail or a single-mode face-to-face design. Only for small available budgets or large allowable error the single-mode mail design seems to perform better. Direct generalization of these conclusions would, however, be inappropriate. Indeed, general conclusions require replications of this study using different survey items, different populations, different statistics, and different cost specifications.

First, the MSE's of the three designs are only compared for one particular question, i.e. a question about survey liking. Moreover, this question probably is very susceptible to selection error and measurement error and this susceptibil-

ity may cause large differences between different survey designs. Other questions, which may be less susceptible to selection and measurement error, might provide less clear differences between different survey designs. When selection effects were smaller, the tipping point between a mixed-mode and single-mode design would be situated at a higher budget because selection error would be less severe in the single-mode design relative to the mixed-mode design. When measurement effects were smaller, in contrast, the tipping point between both designs would be situated at a lower budget because measurement error would be less severe in the mixed-mode design relative to the single-mode design.

Second, the MSE's of the three designs are only compared for one particular statistic, i.e. the mean. The analysis of other, more complex statistics might provide different results. For example, higher-order moments are known to be more sensitive to small sample sizes and better results might be obtained for these statistics by a cheap mail survey compared to the more expensive face-to-face and mixed-mode designs, even for larger budgets.

Third, the MSE's of the three designs are only compared for one particular survey population, i.e. the Flemish population. Survey traditions in other countries might differ from the Flanders case. As such, one survey design might come with completely different selection and measurement error when used in different countries or for different populations. As a result, the three survey designs might compare differently when other populations are considered.

Fourth, the MSE's of the three designs are only compared for one particular cost specification. Indeed, the analysis of the MSE's started from the situation in which the data collection is outsourced to a commercial survey organization, and from a linear relation between budget and gross sample size. In other situations, the fixed and variable costs of all three survey designs might be different, and the relation between budget and gross sample size might be non-linear. In that case, results might differ.

Fifth, the MSE's of the three designs are only compared using one particular analysis model for the estimation of the target mean. Indeed, within this illustration, the front-door model is used including only one single covariate about survey pleasure. The principal question is whether this covariate meets all the required assumptions in order to provide an unbiased estimate of the target mean. Future research might come up with better front-door as well as back-door covariates, and these covariates may provide different conclusions. Another alternative might be to use a single-mode sample as the reference sample. In this illustration, the single-mode face-to-face sample could also be chosen as the reference sample because it has a response rate close to the response rate of the mixed-mode sample. This choice would not require covariate adjustment techniques if face-to-face data-collection is also chosen as the benchmark mode. This strategy, however, would prevent choosing the mail questionnaire as the benchmark mode.

In summary, in order to draw general conclusions about the suitability of a mixed-mode mail—face-to-face survey design relative to its single mode face-to-face and single-mode mail counterparts, the study should be replicated with other target variables, other target statistics, other target populations, and within other financial situations. Further, it can also be noted that no conclusions can be drawn about survey modes other than the mail questionnaire and face-to-face interview and mixed-mode designs other than a sequential design.

4 Discussion

This paper discussed a possible procedure to evaluate the usefulness of mixed-mode survey designs relative to their single-mode counterparts given a specific available survey budget. The Mean Squared Error (MSE), which quantifies the expected error of a design, is used to compare the mixed-mode and single-mode designs. However, this procedure only provides relative comparisons and might possibly be not optimally accurate.

First, the procedure is relative because the MSE requires knowledge about the population, but this knowledge is usually absent. As an alternative, one particular mode is chosen as a benchmark to measure the target variable, and one realized sample is chosen as a reference to represent the population (see assumptions 1 and 2). If a different benchmark mode and/or reference population is chosen, the results will change and may lead to different conclusions. However, such differences can easily be investigated by using sensitivity analyses in which the benchmark mode and reference sample are switched. Nevertheless, such analyses of course do not allow for estimating absolute survey error introduced by different designs. Research to absolute error reduction and estimation procedures are still required in future studies.

Second, the procedure might not be optimally accurate because the benchmark mode and reference sample might not completely overlap. In this paper, the mixed-mode sample is used as the reference population, but this a priori excludes direct estimation of all responses within the benchmark mode. As a solution, the back-door and/or front-door models can be used to correct biased responses from the distorted modes. Nevertheless, both these models require covariates which either explain the selection or the measurement effects between the modes. In general, these requirements reduce to assumptions which are hard to check within applied research. Future survey methodological research might focus on the development of appropriate back-door and front-door covariates. Proper back-door variables must measure, among others, peoples' (in)capacity and (un)willingness to respond in particular modes. As a consequence, para-data about or survey questions asking for mode preferences might be topics for further research so that better back-door covariates can be developed. Proper front-door variables should try to measure causes of measurement error like, among others, response burdens, satisficing, acquiescence, or social desirability. Potential front-door variables might be para-data or survey questions about, among others, survey pleasure, perceived privacy, the number of item non-response, or primacy and recency effects. The operationalisation of such variables might also be topics for further re-

search so that better front-door covariates can be developed.

Further, the advantage of a mixed-mode design depends on several factors like the target variable, the target statistic, the target population, and the particular context in which the data collection is financed. As already mentioned in the illustration, assessing the advantage of mixed-mode surveys requires further replications of this study using different survey items, different populations, different statistics, and different cost specifications. The investigation of all possible survey situations might require some time and research before one can draw unconditional conclusions about the advantage of mixed-mode and single-mode designs.

To end, we also would like to point out one contradiction of the paper. The target mean is estimated from the mixed-mode sample by using back-door or front-door adjustment, but these adjusting covariates are ignored for the estimation of the mixed-mode design-specific mean and variance. If the mixed-mode design-specific mean and variance were calculated conditional on the back- or front-door covariates, the mixed-mode survey would, however, provide unbiased estimates and the MSE of the mixed-mode design would only include variance. Nonetheless, such an analysis model would strongly complicate the mixed-mode MSE decomposition. For that reason, this exercise is omitted from the current paper and left for a follow-up study.

Acknowledgements

References

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(2), 233–255.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social Science Research*, *38*, 1–18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail and mixed-mode surveys: the tailored design method* (3rd ed.). Hoboken: Wiley.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *1*, 151–182.

Gelman, A., Carlin, J. S., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton (Fla.): Chapman and Hall/CRC.

Greenfield, T. K., Midanik, L. T., & Rogers, J. D. (2000). Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*, *95*(2), 277–284.

Hayashi, T. (2007). The possibility of mixed-mode surveys in sociological studies. *International Journal of Japanese sociology*, *16*(1), 51–63.

Heerwegh, D., & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, *27*(1), 49–63.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79–125.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*, 449–484.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3–20.

Little, R. J. A. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546–556.

Little, R. J. A. (2012). Calibrated bayes, an alternative inferential paradigm for official statistics. *Journal of Official Statistics*, *28*(3), 309-334.

Little, R. J. A. (2013). In praise of simplicity not mathamatistry! ten simple powerful ideas for the statistical scientist. *Journal of the American Statistical Association*, *108*(502), 359-369.

Loosveldt, G., & Storms, V. (2008). Measuring public opinions about surveys. *International Journal of Public Opinion Research*, *20*(1), 74–89.

Lugtig, P., Lensvelt-Mulders, G. J. L. M., Frerichs, R., & Greven, A. (2011). Estimating nonresponse bias and mode effects in a mixed-mode survey. *International Journal of Market Research*, *53*(3), 669–686.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs – a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Morgan, S. L., & Winship, C. (2009). *Counterfactuals and causal inference: methods and principles for social research*. New York, (N.Y.): Cambridge university press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). New York (N.Y.): Cambridge University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, *66*(5), 688–701.

Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, *5*(3), 193–212.

Storms, V., & Loosveldt, G. (2005). *Procesevaluatie van het veldwerk van een mixed mode survey naar het surveyklimaat in vlaanderen*. Leuven: KUL, Centrum voor sociologisch onderzoek.

Vannieuwenhuyze, J., & Loosveldt, G. (2013). Evaluating relative mode-effects in mixed mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods and Research*, *42*(1), 82–104.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2014). Evaluating mode effects in mixed-mode survey data using co-

variate adjustment models. *Journal of Official Statistics*, *30*(1), 1–21.

Voogt, R. J., & Saris, W. E. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics*, *21*(3), 367–387.

Weisberg, H. F. (2005). *The total survey error approach: a guide to the new science of survey research*. Chicago (Ill.): University of Chicago.

Weisberg, H. F. (2010). *Bias and causation: Models and judgment for valid comparisons*. Hoboken, New Jersey: Wiley.